



Optimal Checkpointing Strategies for Iterative Applications

Yishu Du, Loris Marchal, Guillaume Pallez, Yves Robert

► To cite this version:

Yishu Du, Loris Marchal, Guillaume Pallez, Yves Robert. Optimal Checkpointing Strategies for Iterative Applications. [Research Report] RR-9371, Inria - Research Centre Grenoble – Rhône-Alpes. 2020. hal-02980455

HAL Id: hal-02980455

<https://inria.hal.science/hal-02980455>

Submitted on 27 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Optimal Checkpointing Strategies for Iterative Applications

Yishu Du, Loris Marchal, Guillaume Pallez, Yves Robert

**RESEARCH
REPORT**

N° 9371

October 2020

Project-Team ROMA



Optimal Checkpointing Strategies for Iterative Applications

Yishu Du^{*†}, Loris Marchal[†], Guillaume Pallez[‡], Yves Robert^{†§}

Project-Team ROMA

Research Report n° 9371 — October 2020 — 37 pages

Abstract: This work provides an optimal checkpointing strategy to protect iterative applications from fail-stop errors. We consider a very general framework, where the application repeats the same execution pattern by executing consecutive iterations, and where each iteration is composed of several tasks. These tasks have different execution lengths and different checkpoint costs. Assume that there are n tasks and that task a_i , where $0 \leq i < n$, has execution time t_i and checkpoint cost C_i . A naive strategy would checkpoint after each task. A strategy inspired by the Young/Daly formula would select the task a_{\min} with smallest checkpoint cost C_{\min} and would checkpoint after every p^{th} instance of that task, leading to a checkpointing period $P_{YD} = pT$ where $T = \sum_{i=0}^{n-1} a_i$ is the time per iteration. One would choose the period so that $P_{YD} = pT \approx \sqrt{2\mu C_{\min}}$ to obey the Young/Daly formula, where μ is the application MTBF. Both the naive and Young/Daly strategies are suboptimal. Our main contribution is to show that the optimal checkpoint strategy is globally periodic, and to design a dynamic programming algorithm that computes the optimal checkpointing pattern. This pattern may well checkpoint many different tasks, and this across many different iterations. We show through simulations, both from synthetic and real-life application scenarios, that the optimal strategy significantly outperforms the naive and Young/Daly strategies.

Key-words: fault-tolerance, checkpoint strategy, fail-stop error, iterative application.

^{*} Tongji University, Shanghai, China

[†] LIP, École Normale Supérieure de Lyon, CNRS & Inria, France

[‡] Inria Bordeaux Sud Ouest, France

[§] University of Tennessee Knoxville, USA

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Stratégies de checkpoint optimales pour les applications itératives

Résumé : Ce rapport propose une stratégie de checkpoint optimale pour protéger les applications itératives des erreurs fatales. On s'intéresse à un cadre général, où l'application répète le même schéma et exécute des itérations consécutives, et où chaque itération est composée de plusieurs tâches. Ces tâches ont des longueurs différentes, et des coûts de checkpoint différents également. Supposons avoir n tâches, et que la tâche a_i , pour $0 \leq i < n$, a un temps d'exécution t_i et un coût de checkpoint C_i . Une stratégie naïve prendrait un checkpoint après chaque tâche. Une stratégie inspirée par la formule de Young/Daly choisirait la tâche a_{\min} dont le coût de checkpoint C_{\min} est minimal, et prendrait un checkpoint après chaque $p^{\text{ème}}$ instance de cette tâche, ce qui conduit à une période de checkpointing $P_{YD} = pT$ où $T = \sum_{i=0}^{n-1} a_i$ est le temps d'une itération. On choisirait alors p pour que $P_{YD} = pT \approx \sqrt{2\mu C_{\min}}$ obéisse à la formule de Young/Daly, où μ est le MTBF (temps moyen entre deux pannes) de l'application. La stratégie naïve et celle de Young/Daly sont toutes deux sous-optimales. Notre contribution principale est de montrer que la stratégie optimale est globalement périodique, et de proposer un algorithme de programmation dynamique qui calcule la période de ce checkpoint optimale. Cette dernière peut prendre le checkpoint de plusieurs tâches, et ce sur plusieurs itérations. Des simulations basées sur des scénarios applicatifs à la fois synthétiques et issus d'applications réelles, montrent bien le gain qu'apporte la stratégie optimale.

Mots-clés : tolérance aux pannes, checkpoint, application itérative.

1 Introduction

Deploying scientific applications at large scale requires fault-tolerant mechanisms. State-of-the-art supercomputers such as Fugaku, Summit or Sunway TaihuLight (respectively ranked 1st, 2nd and 4th in the TOP500 ranking [43]) are now embedding millions of cores. These very large systems are prone to failures: even if each of their cores has a very low probability of failure, the failure probability of the whole system is much higher. More precisely, assume that the *Mean Time Between Failure* (MTBF) of each computing resource is around 10 years, which means that such a resource should experience an error only every ten years on average, and which explains why computing resources are individually very reliable. When running a simulation code on 100,000 of these resources in parallel, the MTBF is reduced to only 50 minutes [24]: on average one node of the computing platform crashes every 50 minutes. With one million of such resources, the MTBF gets as small as five minutes, while codes deployed on such extreme-scale platforms usually last for hours or days. As the demand for computing power increases, failures cannot be ignored anymore, and fault-tolerant mechanisms must be deployed.

The classical way of dealing with failures in extreme-scale computing systems consists of Checkpoint/Rollback mechanisms. A *checkpoint* of the application is taken periodically, that is, the state of the application (usually the whole content of its memory) is written onto reliable storage. Whenever one of the computing resources experiences a failure, the application pauses and restarts from the last valid checkpoint. Several studies have focused on the crucial question of the optimal checkpointing period, defined as the time between two consecutive checkpoints. On the one hand, if checkpoints are taken too often, time is wasted in costly I/O operations. On the other hand, if checkpoints are too infrequent, time will be wasted in recomputing large portions of the application after each failure. Interestingly, reliability was already a question in the early days of computing: in the 70s, Young has proposed a first-order approximation of the optimal time between two checkpoints that minimizes the expected duration of the whole computation [47]. Young's approximation has then been refined by Daly thirty years later [10]. The Young/Daly approach assumes that a checkpoint can be taken anytime during the computation, and that the time needed to take a checkpoint is constant (which corresponds to a constant size of the data to save).

When designing checkpoint/restart strategies for task-based workflows, it is natural to take checkpoint between the completion of some task and the beginning of its successor. This way, the checkpoint mechanism can be provided by the workflow management system without having to modify the code of each task. However, this restricts the time-steps at which checkpoints can be taken and makes the optimization problem of selecting the best checkpoint times more difficult. Furthermore, the data to checkpoint, is

now the output of the tasks and may have different sizes for different tasks of the linear workflow.

In this paper, we focus on designing optimal checkpoint strategies for iterative workflows expressed as *pipelined linear workflows*: we consider workflows made of a large number of iterations, each iteration being a linear chain of tasks. The typical example is an application consisting of an outer loop “While convergence is not met, do”, and where the loop body includes a sequence of large parallel operations. The objective is to find which task outputs should be saved on stable storage in order to minimize the expected duration of the whole computation. To the best of our knowledge, this is an open problem, despite the practical importance and ubiquity of pipelined linear workflows in High-Performance Computing (HPC). Indeed, the simple case of a unique linear chain of tasks (a pipelined linear workflow with a single iteration) has been solved by Toueg et al. [44] using a dynamic programming algorithm. On the contrary, the problem for workflows with general directed graphs has been shown $\#P$ -complete¹ [21]. The study for pipelined linear workflows (a linear chain with several iterations) is challenging, and the main contribution of this paper is to provide a complete answer:

- We prove that there exists an optimal checkpointing strategy which is periodic. It consists in a pattern of task outputs to checkpoint, where this pattern spans over a set of iterations of bounded size. This pattern is repeated over and over throughout the execution.
- We provide a dynamic programming algorithm which is polynomial in the number of operations included in the outer loop to compute the optimal periodic checkpoint pattern. The complexity of the algorithm does not depend on the number of iterations of the outer loop. This pattern may well checkpoint many different tasks, and this across many different iterations.
- We conduct an extensive set of simulations to compare the optimal strategy to two natural competitor strategies: the first one checkpoints after each task while the second one is an extension of the Young/Daly formula for iterative applications. Our simulations with both synthetic and real-life workflow instances demonstrate that our optimal strategy provide significant improvement over the simpler competitors.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 presents a detailed model for the problem and states the objective function. Section 4 outlines the optimal checkpoint strategy. Section 5 reports a comprehensive set of experimental results, based upon both synthetic workflows and on workflows arising from two real-life applications.

¹ $\#P$ -complete problems are at least as hard as NP-complete problems

Finally, Section 6 provides concluding remarks and directions for future work.

2 Related Work

We survey related work in this section. We start with checkpointing in Section 2.1. Then we discuss iterative applications with cyclic tasks in Section 2.2. We end with fault-tolerance methods for iterative applications in Section 2.3.

2.1 Checkpointing

Checkpoint-restart is one of the most used strategy to deal with fail-stop errors, and several variants of this policy have been studied, see [24] for an overview. The natural strategy is to checkpoint periodically, and one must decide how often to checkpoint, i.e., must derive the optimal checkpointing period. An optimal strategy is defined as a strategy that minimizes the expectation of the execution time of the application. For a divisible-load application, given the checkpointing cost C and platform MTBF μ , the classical formula due to Young [47] and Daly [10] states that the optimal checkpointing period is $P_{YD} = \sqrt{2\mu C}$.

Going beyond divisible-load applications, some works have studied linear workflows, i.e., applications that can be expressed as a linear chain of (parallel) tasks. Checkpointing is only possible right after the completion of a task, and the problem is to determine which tasks should be checkpointed. This problem has been solved by Toueg and Babaoglu [44] using a dynamic programming algorithm. We stress that this latter approach is not suited to iterative applications. Indeed, consider an iterative application with a large number of iterations, say $N_{iter} = 10,000$ iterations, and assume that $n = 10$ (10 tasks per iteration). One solution to find an optimal checkpointing strategy could be: (i) unroll the loop and build a linear chain of $n \times N_{iter} = 100,000$ tasks; (ii) apply the algorithm of [44] to this huge chain and return the optimal solution. However, the cost of this algorithm is quadratic in the value of N_{iter} . Worse, if we re-execute the same application for $N_{iter} = 20,000$ iterations, we have to recompute the optimal solution from scratch. On the contrary, our approach provides a generic and compact solution that does not depend upon the value of N_{iter} .

Recently, the results of [44] have been extended to deal with linear chains whose tasks do not have constant execution times but instead obey some probability distribution [13]. As pointed out above, for general workflows, deciding which tasks to checkpoint has been shown $\#P$ -complete [21], but the results of [3] show that if the graph is scheduled in a sequential manner (linearized), then one can derive an optimal checkpointing strategy. In this

paper, we focus on pipelined linear workflows, i.e., on applications expressed as a linear chain of tasks that repeats iteratively.

2.2 Iterative applications

Iterative methods are popular for solving large sparse linear systems, which have a wide range of applications in several scientific and industrial problems. There are many classic iterative methods including stationary iterative methods like the Jacobi method [36], the Gauss-Seidel method [36] and the Successive Overrelaxation method (SOR) [16, 45], and non-stationary iterative methods like Krylov subspace methods, including Generalized Minimal Residual method (GMRES) [37], Bi-conjugate Gradient Stabilized method (BiCGSTAB) [20], Generalized Conjugate Residual method (GCR) [14], together with their ABFT (algorithm-based fault-tolerance) variants [1, 2, 28].

Krylov subspace methods fit perfectly our model: the outer iteration corresponds to an iteration of the application in our model, while the inner iterations in each increasing Krylov subspace correspond to the tasks in our model. See Algorithm 2 in Section 5.1 for an illustration of GCR, with an inner loop of n tasks, each incurring an increasing checkpoint cost. We report experiments for GCR in Section 5.2.

The class of iterative applications goes well beyond sparse linear solvers. Uncertainty Quantification (UQ) workflows explore a parameter space in an iterative fashion [29, 35]. This class also encompasses many image and video processing software which operate a chain of computations kernels (each being a task) on a sequence of data sets (each corresponding to an iteration). Examples include image analysis [39], video processing [19], motion detection [27], signal processing [9, 23], databases [6], molecular biology [34], medical imaging [18], and various scientific data analyses, including particle physics [11], earthquake [26], weather and environmental data analyses [34].

2.3 Fault-tolerance methods for iterative applications

The literature devoted to the study of fault-tolerance methods for iterative linear solvers can be divided into two categories, depending upon whether the focus is on soft errors or on fail-stop errors.

There are some works dealing with soft errors. Chen presented online-ABFT in [8], a technique that can detect soft errors in the specific Krylov subspace iterative methods by leveraging the orthogonality relationship of two vectors in the middle of the program execution. For general iterative methods, Tao et al. [42] presented a new online-ABFT approach to detect and recover soft errors by combining a novel checksum-based encoding scheme with a checkpoint/rollback scheme. According to the specific properties of GMRES algorithm, Bridges et al. [5] and Elliott et al. [15] proposed the FT-GMRES algorithm using selective reliability. Similarly,

Sao and Vuduc [38] proposed the self-stabilizing conjugate gradient method (CG) in view of the special properties of CG algorithm: they check that orthogonality is preserved by recomputing scalar products that should be zero and restarting whenever a threshold is exceeded. Ozturk et al. [30] proposed a decreasing energy norm based on the mathematical properties to detect soft errors leading to silent data corruption (SDC) for GMRES, CG and Conjugate Residual method (CR).

There are some works dealing with fail-stop errors. To reduce the fault tolerance overhead incurred by checkpointing, Chen [7] proposed a recovery method for iterative methods without checkpointing based on the specific properties of iterative methods. Tao et al. [41] improved the checkpointing performance for iterative methods under a novel lossy checkpointing scheme. Langou et al. [28] presented a lossy approach which is a checkpoint-free fault tolerant scheme for parallel iterative methods. The iterative method is restarted with a new vector which is a new approximate solution recovered from a fail-stop error by using the data of the non-failed processors. Agullo et al. [1, 2] extended this approach by computing a well-suited initial guess which is defined by interpolating the lost entries of the current iterate vector available on surviving nodes, in order to restart the Krylov method. Pachajoa et al. [31] compared exact state reconstruction (ESR) approach based on the method proposed by Chen [7] with the heuristic linear interpolation (LI) approach by Langou et al. [28] and Agullo et al. [1, 2]. They later extended the ESR approach for protecting the PCG method against multiple and simultaneous node failures [32, 33]. Altogether, fault-tolerance methods proposed to mitigate the impact of fail-stop errors in iterative applications are application-specific, and can only be applied to a particular class of iterative algorithms. Moreover, their performance highly depends upon specific properties of the algorithms, and, for instance, considerably vary from one Krylov method to another.

To the best of our knowledge, this work is the first general-purpose approach to deal with fail-stop errors in iterative applications. Our optimal checkpointing strategy is agnostic of any specific property of the target iterative application. Instead, it abstracts the iterative application as a chain of cyclic tasks, and provides the optimal periodic checkpoint pattern based only upon generic information such as task durations and checkpoint costs.

3 Model

In this section, we detail the application and platform models. Then we define checkpoint strategies and formally state the optimization objective.

3.1 Application model

We consider an iterative application \mathcal{A} . Each iteration of the application consists of n tasks a_i , where $0 \leq i < n$, task a_i has length t_i and memory footprint M_i . We define the length of an iteration as $T = \sum_{i=0}^{n-1} t_i$.

The tasks are executed consecutively: let $i[n]$ denote the remainder of the integer division of i by n (modulo operation); then a task a_i is always followed by a task $a_{i+1[n]}$. We assume that the application executes for a long time and consider an unbounded number of iterations. For short we write $\mathcal{A} = (a_0, \dots, a_{n-1})^\infty$. As stated before, we execute tasks one after the other. The first executed task is a_0 , followed by a_1 , and so on. The n^{th} task is a_{n-1} and the $(n+1)$ st task is a_0 again. In general, the k^{th} task is $a_{k-1[n]}$. Note that we index tasks from 0 to use the modulo operation, hence this shift when counting executed tasks.

We assume that the tasks of the application can be checkpointed at the end of their execution. We consider a general model where the checkpoint time of task a_i is c_i and its recovery time is r_i . We refer to c_i and r_i as operations of type i . We do not assume that $c_i = r_i$; instead, we simply assume monotone I/O costs:

$$\text{for all } i, j, c_i \geq c_j \implies r_i \geq r_j. \quad (1)$$

Essentially this assumption states that if a task is longer to checkpoint than another one, then restarting from this checkpoint is also longer. This is coherent with the fact that checkpoint and recovery costs are often closely related, and are a function of the volume of data to save. Furthermore, this assumption is general enough to account for different read and write bandwidths. In the experiments, we further assume that the checkpoint and recovery times c_i and r_i linearly depend upon the memory footprint M_i , but this is only one way to instantiate the model, and all the theoretical results hold for the general model with arbitrary monotone values of c_i and r_i .

3.2 Platform model

We consider a parallel platform subject to fail-stop failures. We assume that the iterative application experiences failures whose inter-arrival times follow an Exponential distribution $Exp(\lambda)$ of parameter $\lambda > 0$, whose PDF (Probability Density Function) is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. The MTBF is $\mu = \frac{1}{\lambda}$ and corresponds to the MTBF of individual processors divided by the total number of processors enrolled in the application [24]. When hit by a failure, the platform is unavailable during a downtime D , which is the time to replace the faulty processor by a spare.

We can always scale the length of the tasks by an appropriate factor so that the platform has unit speed, meaning we can speak of a task length, or duration, or work amount, indifferently.

3.3 Schedule

Informally, a schedule defines which tasks are checkpointed. A priori, there is no reason for a schedule to enforce a regular pattern of checkpoints that repeats over time. In other words, a schedule can be aperiodic. However, one major contribution of this work is to show that periodic schedules are optimal, and to exhibit the optimal period as the output of a polynomial-time algorithm. We need a few definitions before stating the objective function to be minimized by *optimal* schedules. First we identify a schedule with the list of the tasks that it checkpoints:

Definition 1 (Schedule). A schedule \mathcal{S} is an infinite increasing sequence $\mathcal{S} = (m_1, m_2, \dots)$ which represents the list of checkpointed tasks: the m_i^{th} task (i.e., task number m_i) is checkpointed, and the tasks whose number does not belong to the list are not checkpointed.

In other words, checkpoint number i in the schedule takes place at the end of task number m_i . The cost to checkpoint that task m_i is $c_{m_i-1[n]}$ (because of the index shift noted above). Without loss of generality, we assume that the schedule checkpoints infinitely many tasks, i.e., $\lim_{i \rightarrow \infty} m_i = \infty$. Indeed, consider any task in the application: eventually there must be a checkpoint after that task, otherwise the expected execution time from that task on is not bounded, because the fault-rate λ is nonzero.

A schedule \mathcal{S} can be viewed as a succession of task chunks between two consecutive checkpoints. We use the following notations for the i^{th} chunk between checkpoint number $i-1$ (or the beginning of the execution if $i=1$) and checkpoint number i :

- The length of the tasks in the chunk is

$$W_i^{\mathcal{S}} = \sum_{j=m_{i-1}+1}^{m_i} t_{j-1[n]}$$

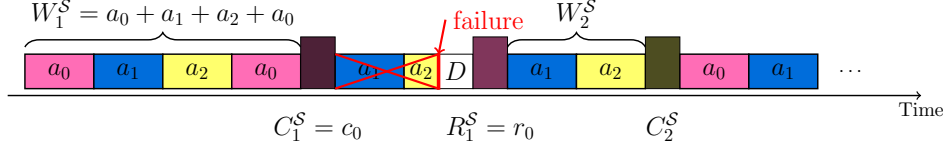
- The checkpoint cost at the end of the chunk is the cost of checkpoint number i , namely

$$C_i^{\mathcal{S}} = c_{m_i-1[n]}$$

- The recovery cost when re-executing the chunk is the cost of recovering from checkpoint number $i-1$, namely

$$R_{i-1}^{\mathcal{S}} = r_{m_{i-1}-1[n]}$$

When $i=1$ (first chunk), we let $m_0=0$, and $R_{-1}^{\mathcal{S}}$ denotes the cost of reading input data.

Figure 1: Notations drawn in a schedule for an application with $n = 3$ tasks.

3.4 Objective function

Intuitively, a good schedule will minimize the slowdown during the execution. This slowdown comes from two sources of overhead: the checkpoints that are inserted, and the time lost due to failures. When a failure strikes during execution, the work executed since the last checkpoint is lost; there is a downtime, followed by a recovery, and then the re-execution of the work that has been lost due to the failure. Altogether, the overhead is not deterministic and varies from one execution to the other, hence we aim at minimizing the slowdown in expectation.

Given a schedule \mathcal{S} , we rely on a well-known formula to compute the expected execution time of a chunk. Indeed, the expected execution time $\mathbb{E}_\lambda(w, c, r)$ to execute w consecutive seconds of work followed by a checkpoint of size c with a recovery of size r is given by [24]

$$\mathbb{E}_\lambda(w, c, r) = \left(\frac{1}{\lambda} + D \right) e^{\lambda r} \left(e^{\lambda(w+c)} - 1 \right) \quad (2)$$

The expected time to execute the chunk number i is thus $\mathbb{E}_\lambda(W_i^S, C_i^S, R_{i-1}^S)$. Hence the expected time to execute the first i chunks is $\sum_{j=1}^i \mathbb{E}_\lambda(W_j^S, C_j^S, R_{j-1}^S)$. The slowdown incurred for the first i chunks (i.e., up to checkpoint number i) is therefore

$$\text{SD}_i(\mathcal{S}) = \frac{\sum_{j=1}^i \mathbb{E}_\lambda(W_j^S, C_j^S, R_{j-1}^S)}{\sum_{j=1}^{m_i} t_{j-1}[n]} \quad (3)$$

In Equation (3), the numerator is the expected time to execute the first i chunks, while the denominator is the duration of the tasks up to checkpoint number i , which corresponds to the resilience-free and failure-free execution.

Unfortunately, there is no reason that $\lim_{i \rightarrow \infty} \text{SD}_i(\mathcal{S})$ would exist. However, we can use the upper limit of $\text{SD}_i(\mathcal{S})$ to define the slowdown of schedule \mathcal{S} :

Definition 2 (Slowdown). The slowdown SD of a schedule is:

$$\text{SD}(\mathcal{S}) = \overline{\lim}_{i \rightarrow \infty} \text{SD}_i(\mathcal{S}) \quad (4)$$

We know that this upper limit is bounded for some schedules. Consider for instance the schedule \mathcal{S} that checkpoint all tasks: $m_i = i$ for all $i \geq 1$.

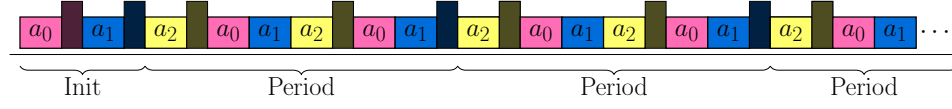


Figure 2: A periodic schedule where the period is repeated over time.

This schedule repeats the same pattern of checkpoints every iteration, so that its slowdown is

$$\text{SD}(\mathcal{S}) = \frac{\text{SD}_n(\mathcal{S})}{T} = \frac{\sum_{j=0}^{n-1} \left(\frac{1}{\lambda} + D \right) e^{\lambda r_{j-1}[n]} (e^{\lambda(t_j + c_j)} - 1)}{T}$$

Recall that T is the length of an iteration. We are now ready to define an optimal schedule:

Definition 3 (Optimal schedule). A schedule is optimal if its slowdown $\text{SD}(\mathcal{S})$ is minimal over all possible schedules.

Note that the definition does not assume that there exists a unique optimal schedule. A major contribution of this paper is to show that there exists an optimal schedule which is periodic, i.e., which repeats the same pattern of checkpoints after some point (see below for the formal definition). This important result will allow us to consider only a finite number of candidate schedules, and to design a polynomial-time algorithm to find an optimal schedule.

3.5 Periodic schedules

Periodic schedules are natural schedules that can be expressed in a compact form. As already mentioned, after some possible initialization phase, a periodic schedule repeats the same sequence of checkpoints over and over. Here is the formal definition:

Definition 4 (Periodic schedules). A schedule (m_1, m_2, \dots) is periodic if there exists two indices i_0 and k_0 such that for all $i > i_0$, $m_i - m_{i-1} = m_{i+k_0} - m_{i+k_0-1}$.

These schedules are illustrated in Figure 2. Intuitively, the schedule enters its steady state after checkpoint number i_0 (with possibly $i_0 = 0$): the period starts right after task number m_{i_0} , and then repeats the same sequence of k_0 checkpoints: the first checkpoint of the period is taken after $m_{i_0+1} - m_{i_0}$ tasks, the second one after $m_{i_0+2} - m_{i_0+1}$ tasks, until the last checkpoint of the period, that of task number $m_{i_0+k_0}$. Then the period repeats indefinitely.

For a periodic schedule, the limit $\lim_{i \rightarrow \infty} \text{SD}_i(\mathcal{S})$ always exists, and is given by the slowdown incurred during each (infinitely repeating) period.

Specifically, given i_0 and k_0 in Definition 4, we see from Equation (3) that this slowdown becomes:

$$\frac{\sum_{i=i_0+1}^{i_0+k_0} \mathbb{E}_\lambda(W_i^S, C_i^S, R_{i-1}^S)}{\sum_{j=m_{i_0}+1}^{m_{i_0}+k_0} t_{j-1}[n]}$$

We prove this result formally below. The major results of this work are the following two theorems, which we prove in Section 4 below.

Theorem 1. *There exists a periodic schedule that is optimal.*

Theorem 2. *We can compute an optimal periodic schedule in polynomial time.*

4 Optimal checkpoint strategy

In this section, we present several theoretical results and prove Theorems 1 and 2. Specifically, we start by showing that we can indeed focus on periodic algorithms (Theorem 1) in Section 4.3. Then in Section 4.4, we show that we can compute an optimal periodic schedule in polynomial time. (Theorem 2).

Beforehand, we introduce the definition of a *pattern* which is at the heart of periodic algorithms (Section 4.1), and we present several important properties of patterns in Section 4.2).

4.1 Paths and patterns

Definition 5 (Checkpoint Paths). A *Checkpoint Path* ($P = (i_0, [m_1, \dots, m_{k_P}])$) is a sequence of m_{k_P} tasks $b_0, \dots, b_{m_{k_P}-1}$ such that:

1. for $0 \leq i \leq m_{k_P} - 1$, $b_i = a_{i_0+i[n]}$;
2. for $1 \leq j \leq k_P$, b_{m_j-1} is checkpointed.

Thus the path starts at task $b_0 = a_{i_0}$ and includes m_{k_P} tasks, up to task $b_{m_{k_P}-1} = a_{i_0+m_{k_P}-1[n]}$. The path includes k_P checkpoints, including the checkpoint of its last task. The m_i^{th} task of the pattern is checkpointed, for $1 \leq i \leq k_P$. See Figure 3 for an illustration.

We use the following notations for a path P : for $1 \leq i \leq k_P$, we define $W_i^P = \sum_{j=m_{i-1}}^{m_i-1} t_{i_0+j[n]}$ (with the special case $m_0 = 0$), $C_i^P = c_{i_0+m_i-1[n]}$, $R_i^P = r_{i_0+m_i-1[n]}$ (with the special case: $R_0^P = r_{i_0-1[n]}$). We define the length (ℓ) of a checkpoint path $\ell(P) = \sum_{i=1}^{k_P} W_i^P$ and its expected execution

time, or cost (\mathcal{C}):

$$\begin{aligned}\mathcal{C}(P) &= \sum_{i=1}^{k_P} \mathbb{E}_\lambda (W_i^P, C_i^P, R_{i-1}^P) \\ &= \left(\frac{1}{\lambda} + D \right) \sum_{i=1}^{k_P} e^{\lambda R_{i-1}^P} \left(e^{\lambda(W_i^P + C_i^P)} - 1 \right)\end{aligned}$$

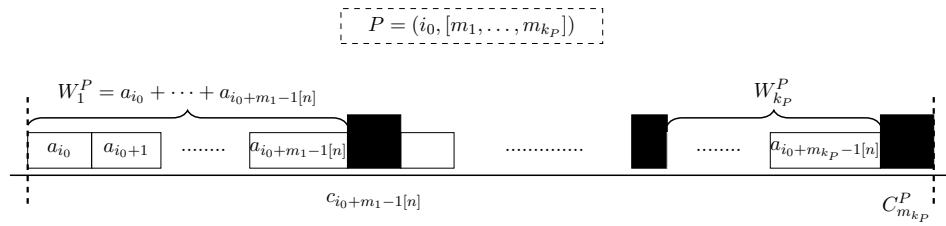


Figure 3: The sequence of operations of a checkpoint Path $P = (i_0, [m_1, \dots, m_{k_P}])$ (Def. 5). Its length is the sum of its *useful* work (white boxes). Its cost correspond to its expected execution time if a checkpoint was taken right before its start.

Definition 6 (Patterns). A *Checkpoint Pattern* is a checkpoint path $P = (i_0, [m_1, \dots, m_{k_P}])$ such that $m_{k_P} = 0[n]$. Note, for pattern P , its length is $\ell(P) = \frac{m_{k_P}}{n}T$, where $T = \sum_{i=0}^{n-1} t_i$.

Such patterns are basic blocks to define periodic schedules. We detail this relation below in Section 4.3.

Definition 7 (Slowdown of a path (or pattern)). The slowdown of a path is defined as:

$$S(P) = \frac{\mathcal{C}(P)}{\ell(P)}$$

4.2 Pattern properties

Using these definitions, we show the following result:

Theorem 3. *Given an schedule \mathcal{S} of slowdown $SD(\mathcal{S})$, there exists a pattern \mathcal{P} such that $S(\mathcal{P}) \leq SD(\mathcal{S})$.*

Following the proof: Theorem 3 aims at showing the existence of a pattern whose slowdown is at most that of any algorithm (hence including optimal algorithms). To do so, we construct a sequence of patterns whose slowdown converges to the requested slowdown (Lemma 1)



Figure 4: Partitioning the schedule into patterns around the checkpoints c_2 (yellow).

based on sequences taken from the algorithm \mathcal{S} . In addition, we impose that this sequence satisfies a *size property*, i.e. that each element of this sequence contains at most n checkpoints (Corollary 1). This then helps to find a pattern whose slowdown is exactly that of \mathcal{S} .

We start by proving a series of results.

Lemma 1. *Given a schedule \mathcal{S} , there exists a sequence of patterns $(\mathcal{P}_r)_r$ such that, for all $r \in \mathbb{N}$, $S(\mathcal{P}_r) \leq SD(\mathcal{S}) + 1/r$.*

Proof. Consider a schedule $\mathcal{S} = (m_i)_{i \in \mathbb{N}}$ of finite slowdown $SD(\mathcal{S})$. There exists a checkpoint type i_0 which is taken an infinite number of times. We denote by σ_{i_0} the function such that, for all i , $\sigma_{i_0}(i)$ is the i^{th} occurrence of checkpoint c_{i_0} in the schedule \mathcal{S} (we set $\sigma_{i_0}(0) = -1$).

In the following we partition the schedule into paths:

$$\begin{cases} \mathcal{M}_1 = (0, [m_1, \dots, m_{\sigma_{i_0}(1)}]) \\ \mathcal{M}_i = (i_0 + 1, [(m_{\sigma_{i_0}(i-1)+1} - m_{\sigma_{i_0}(i-1)}), \dots, \\ (m_{\sigma_{i_0}(i)} - m_{\sigma_{i_0}(i-1)})]) \end{cases} \quad (\forall i > 1)$$

Intuitively, \mathcal{M}_1 is the beginning of the schedule until the first checkpoint of type i_0 . Then \mathcal{M}_2 is the pattern starting right after and extending up to the second checkpoint of type i_0 , and so on. In the definition of \mathcal{M}_i , checkpoint indices are shifted to account for the location where the path starts. See Figure 4 for an illustration. By construction, each \mathcal{M}_i is indeed a pattern, except for \mathcal{M}_1 , which is only a path if $i_0 \neq n - 1$.

We now study the slowdown $SD_{\sigma_{i_0}(i)}$ up to the i^{th} checkpoint of type i_0 , i.e., the slowdown of the first i segments. We have

$$\begin{aligned} SD_{\sigma_{i_0}(i)} &= \frac{\sum_{k=1}^i \sum_{j=\sigma_{i_0}(k-1)+1}^{\sigma_{i_0}(k)} \mathbb{E}_\lambda(W_j^S, C_j^S, R_{j-1}^S)}{\sum_{k=1}^i \sum_{j=\sigma_{i_0}(k-1)+1}^{\sigma_{i_0}(k)} t_{j-1}[n]} \\ &= \frac{\sum_{k=1}^i \mathcal{C}(\mathcal{M}_k)}{\sum_{k=1}^i \ell(\mathcal{M}_k)} = \sum_{k=1}^i \alpha_{i,k} \cdot S(\mathcal{M}_k) \end{aligned} \quad (5)$$

where $\alpha_{i,k} = \frac{\ell(\mathcal{M}_k)}{\sum_{j=1}^i \ell(\mathcal{M}_j)}$. Hence $\sum_{k=1}^i \alpha_{i,k} = 1$, and we have expressed $SD_{\sigma_{i_0}(i)}$ as a weighted average of the path slowdowns $S(\mathcal{M}_k)$.

By definition of $\overline{\lim}$ we have:

- $\overline{\lim_{i \rightarrow \infty}} \text{SD}_{\sigma_{i_0}(i)}(\mathcal{S}) \leq \overline{\lim_{i \rightarrow \infty}} \text{SD}_i(\mathcal{S}) = \text{SD}(\mathcal{S});$
- For all r , there exists i_r such that $\forall i > i_r,$

$$\begin{aligned} \text{SD}_{\sigma_{i_0}(i_r)}(\mathcal{S}) &\leq \overline{\lim_{i \rightarrow \infty}} \text{SD}_{\sigma_{i_0}(i)}(\mathcal{S}) + \frac{1}{r} \\ &\leq \text{SD}(\mathcal{S}) + \frac{1}{r}. \end{aligned}$$

Using Equation (5), we obtain:

$$\sum_{k=1}^i \alpha_{i,k} S(\mathcal{M}_k) \leq \text{SD}(\mathcal{S}) + \frac{1}{r}$$

Since this is a weighted average, it means that there exists k_r , where $1 \leq k_r \leq i$ such that $S(\mathcal{M}_{j_r}) \leq \text{SD}(\mathcal{S}) + \frac{1}{r}$. If $k_r \neq 1$, or if $k_r = 1$ and $i_0 = n - 1$, we have found the desired pattern by letting $\mathcal{P}_r = \mathcal{M}_{k_r}$. Otherwise, we redo the same proof using the truncated schedule $\tilde{\mathcal{S}}$ where we delete the first i_0 tasks. Then $\tilde{\mathcal{S}}$ is a valid schedule for a rotation of the original application, namely for the application $\tilde{\mathcal{A}} = (a_{i_0+1[n]}, \dots, a_{i_0})^\infty$, and it has same slowdown as \mathcal{S} . The path $\tilde{\mathcal{M}}_i$ of $\tilde{\mathcal{S}}$ is the same as the path \mathcal{M}_{i+1} of \mathcal{S} for all i , hence all the paths of $\tilde{\mathcal{S}}$ are patterns. We then derive the result just as above. \square

Lemma 2. *For all pattern \mathcal{P} , there exists a pattern $\tilde{\mathcal{P}}$ such that:*

1. $S(\tilde{\mathcal{P}}) \leq S(\mathcal{P});$
2. $\tilde{\mathcal{P}}$ contains at most n checkpoints.

Proof. We show this result by induction on the number of checkpoints in \mathcal{P} . Assume $\mathcal{P} = (i_0, [m_1, \dots, m_k])$, with $k > n$. Then there exists $i_1 < i_2$ such that: $i_0 + m_{i_1}[n] = i_0 + m_{i_2}[n]$ (i.e. the $m_{i_1}^{\text{th}}$ and $m_{i_2}^{\text{th}}$ tasks of the pattern are identical and equal to $a_{i_0+m_1-1[n]}$).

We now consider the two patterns:

$$\begin{aligned} \mathcal{P}_1 &= (i_0 + m_{i_1}[n], [m_{i_1+1} - m_{i_1}, \dots, m_{i_2} - m_{i_1}]); \\ \mathcal{P}_2 &= (i_0 + m_{i_2}[n], [m_{i_2+1} - m_{i_2}, \dots, m_k - m_{i_2}, \\ &\quad m_1 + (m_k - m_{i_2}), \dots, m_{i_1} + (m_k - m_{i_2})]) \end{aligned}$$

Here, we have decomposed the original pattern into three paths, $\mathcal{P}_{\text{begin}}$ (up to the $m_{i_1}^{\text{th}}$ task), \mathcal{P}_1 (from the next task up to the $m_{i_2}^{\text{th}}$ task) and \mathcal{P}_{end} (from the next task up to the end of the pattern). Now, \mathcal{P}_2 is simply the concatenation of \mathcal{P}_{end} followed by $\mathcal{P}_{\text{begin}}$. See Figure 5 for an illustration.

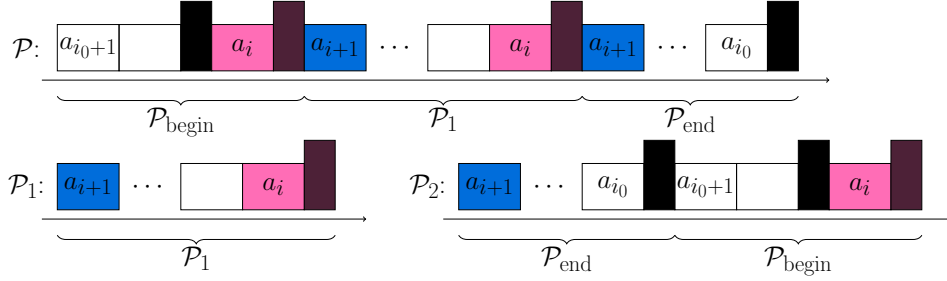


Figure 5: From a pattern \mathcal{P} with two identical checkpoints, c_i , to its decomposition into \mathcal{P}_1 and \mathcal{P}_2 .

We immediately have:

$$S(\mathcal{P}) = \frac{\ell(\mathcal{P}_1)}{\ell(\mathcal{P}_1) + \ell(\mathcal{P}_2)} S(\mathcal{P}_1) + \frac{\ell(\mathcal{P}_2)}{\ell(\mathcal{P}_1) + \ell(\mathcal{P}_2)} \cdot S(\mathcal{P}_2)$$

Because this is a weighted average, then $\min(S(\mathcal{P}_1), S(\mathcal{P}_2)) \leq S(\mathcal{P})$. Each of these patterns have one fewer checkpoint than the initial pattern, which concludes the proof. \square

Corollary 1. *Given a schedule \mathcal{S} , there exists a sequence of patterns $(\tilde{\mathcal{P}}_r)$ for all $r \geq 1$ such that:*

1. *For all r , $\tilde{\mathcal{P}}_r$ contains at most n checkpoints;*
2. *$S(\tilde{\mathcal{P}}_r) \leq SD(\mathcal{S}) + 1/r$.*

This corollary is a direct consequence from Lemma 1, for the existence of a sequence that satisfies the slowdown constraint, and from Lemma 2, for transforming this sequence into a sequence of patterns that include at most n checkpoints.

Following the proof: At this point, we have constructed a sequence of patterns, whose slowdown converges towards $SD(\mathcal{S})$. It remains to show the existence of a pattern that reaches the limit. In order to do so, we show that if the number of checkpoints in a pattern is bounded, then the length of the pattern has to be bounded too, otherwise its slowdown would diverge. This is the result shown in Lemma 3.

Lemma 3. *Given M and k , if \mathcal{P} is a pattern with at most k checkpoints and $S(\mathcal{P}) \leq M$, then there exists a constant $W_{M,k}$ such that $\ell(\mathcal{P}) \leq W_{M,k}$.*

Proof. Given a pattern \mathcal{P} with k checkpoints, and of length $\ell(\mathcal{P}) = W$, we let $W_1^{\mathcal{P}}, W_2^{\mathcal{P}}, \dots, W_k^{\mathcal{P}}$ denote the work between its checkpoints. By definition, $\sum_{i=1}^k W_i^{\mathcal{P}} = \ell(\mathcal{P})$. Hence there exists i_1 such that $W_{i_1}^{\mathcal{P}} \geq \frac{1}{k} \ell(\mathcal{P})$.

We are now interested in the slowdown of the pattern:

$$\begin{aligned}
S(\mathcal{P}) &= \frac{\sum_{i=1}^k \mathbb{E}_\lambda(W_i^{\mathcal{P}}, C_i^{\mathcal{P}}, R_{i-1}^{\mathcal{P}})}{\ell(\mathcal{P})} \\
&\geq \frac{\mathbb{E}_\lambda(W_{i_1}^{\mathcal{P}}, C_{i_1}^{\mathcal{P}}, R_{i_1-1}^{\mathcal{P}})}{\ell(\mathcal{P})} \\
&\geq \frac{\mathbb{E}_\lambda(W_{i_1}^{\mathcal{P}}, 0, 0)}{\ell(\mathcal{P})} = \left(\frac{1}{\lambda} + D\right) \frac{e^{\lambda W_{i_1}^{\mathcal{P}}} - 1}{\ell(\mathcal{P})} \\
&\geq \left(\frac{1}{\lambda} + D\right) \frac{e^{\frac{\lambda}{k} \ell(\mathcal{P})} - 1}{\ell(\mathcal{P})}
\end{aligned}$$

But $\frac{e^{\frac{\lambda}{k}x} - 1}{x}$ tends to infinity when x tends to infinity; hence, because $S(\mathcal{P}) \leq M$, we have that $\ell(\mathcal{P})$ is bounded by a function of M and k . Hence the result. \square

Proof of Theorem 3. We now conclude the proof of Theorem 3. From Corollary 1, we have a sequence of patterns $(\tilde{\mathcal{P}}_r)_r$ with at most n checkpoints and of slowdown $S(\tilde{\mathcal{P}}_r) \leq \text{SD}(\mathcal{S}) + 1/r \leq 2\text{SD}(\mathcal{S})$.

From Lemma 3, there exists an upper bound such that, for all r , $\ell(\tilde{\mathcal{P}}_r) \leq \tilde{W}$. We show that there are only a bounded number of patterns that satisfy this property:

- Since the length of a pattern is a multiple of $T = \sum_{i=0}^{n-1} t_i$, there are at most $K = \lfloor \frac{\tilde{W}}{T} \rfloor$ possible lengths.
- For a length kT , $1 \leq k \leq K$, there are kn possible checkpoint locations and at most n checkpoints, hence at most $\binom{kn}{n}$ patterns.

Hence in total the number of possible patterns is upper bounded by $K \binom{Kn}{n}$. The set $\{S(\tilde{\mathcal{P}}_r) | r \geq 1\}$ is finite and admits a minimum S_{\min} . Let r_0 one index achieving the minimum: $S_{\min} = S(\tilde{\mathcal{P}}_{r_0})$.

Finally, we show that $S_{\min} \leq \text{SD}(\mathcal{S})$: indeed, otherwise there would exist r such that $S_{\min} > \text{SD}(\mathcal{S}) + \frac{1}{r}$ and we would have $S(\tilde{\mathcal{P}}_{r_0}) > S(\tilde{\mathcal{P}}_r)$, thereby contradicting the minimality. Hence the result. \square

4.3 Periodic schedules

Using the properties of patterns, we are ready to derive Theorem 1. We start by rewriting the definition of periodic schedules using patterns. Indeed, the values i_0 and k_0 from Definition 4 allow us to define a pattern that is repeated all throughout the execution. We then select the pattern of minimal length that occurs as early as possible:

Definition 8 (Pattern of a periodic schedule). Given a periodic schedule $\mathcal{S} = (m_1, m_2, \dots)$. Let (k_0, i_0) be the smallest pair (for the lexicographic order) that satisfies: for all $i > i_0$, $m_i - m_{i-1} = m_{i+k_0} - m_{i+k_0-1}$, and $m_{i_0+k_0} - m_{i_0} = 0[n]$. We say that

$$P_{\mathcal{S}} = (m_{i_0}[n], [m_{i_0+1} - m_{i_0}, \dots, m_{i_0+k_0} - m_{i_0}])$$

is the *pattern of the schedule*.

The lexicographic order means that we select first a pattern of minimal length, and in case of a tie, the pattern that starts as early as possible.

Theorem 4 (Slowdown of a periodic schedule). *Given a periodic schedule \mathcal{S} , its slowdown is equal to the slowdown of its pattern.*

Proof. Given a periodic schedule \mathcal{S} , let

$$P_{\mathcal{S}} = (m_{i_0}[n], [m_{i_0+1} - m_{i_0}, \dots, m_{i_0+k_0} - m_{i_0}])$$

be its pattern. We study the function $\text{SD}_i(\mathcal{S})$ by decomposing the schedule up to its i^{th} checkpoint into three parts: a first part, up to the beginning of the pattern, i.e. up to checkpoint number i_0 , then a number $k = \lfloor \frac{i-m_{i_0}}{k_0} \rfloor$ of repeating patterns, then a final part (whose length is smaller than $\ell(P_{\mathcal{S}})$). The first and final part become negligible as i tends to infinity, hence

$$\text{SD}(\mathcal{S}) = \overline{\lim_{i \rightarrow \infty}} \text{SD}_i(\mathcal{S}) = \frac{\mathcal{C}(P_{\mathcal{S}})}{\ell(P_{\mathcal{S}})} = S(P_{\mathcal{S}})$$

□

Finally, putting everything together, we obtain the final result:

Proof of Theorem 1. Theorem 3 states that there exists a pattern P whose slowdown is smaller or equal to that of an optimal schedule. In addition, Theorem 4 states that a periodic schedule whose pattern is P has a slowdown equal to that of P , hence it is optimal. □

4.4 Finding the optimal pattern

In this section, we show how to compute the pattern of an optimal periodic algorithm. In the following, we say that a pattern $\mathcal{P} = (i_0, [m_1, m_2, \dots, m_{k_0}])$ is an *optimal pattern*, if it has minimal slowdown.

4.4.1 Bounding the length

Following the proof: In Lemma 3, we have shown that it was possible to bound the length of an optimal pattern, which was helpful to prove the existence of an optimal pattern. In order to derive an optimal solution, we want to use a dynamic program whose complexity depends on the number of tasks in a pattern. Unfortunately, the previous bound may lead to a number of tasks in the pattern which is not polynomially bounded. We now show how one can get a tighter bound (Theorem 5). At the end of this section, we discuss the size of this bound as a function of the problem instance.

In this section we make intensive use of the following *slowdown function*:

$$f(w, c, r) = \frac{\mathbb{E}_\lambda(w, c, r)}{w} = \left(\frac{1}{\lambda} + D \right) e^{\lambda r} \left(\frac{e^{\lambda(w+c)} - 1}{w} \right) \quad (6)$$

Note that we implicitly used the slowdown function when we defined the slowdown of a schedule. We have the following properties:

Lemma 4. *We have the following properties of the slowdown function:*

1. $w \mapsto f(w, c, r)$ has a unique minimum w_c , is decreasing in the interval $[0, w_c]$ and is increasing in the interval $[w_c, \infty)$
2. $c \mapsto f(w, c, r)$ (resp. $r \mapsto f(w, c, r)$) are increasing functions of c (resp. r).

Proof. 2) is obvious. 1) is the result of [4, Theorem 1]. Note that a first-order approximation of w_c is the well-known Young/Daly formula $w_c = \sqrt{\frac{2c}{\lambda}}$ [10, 47]. \square

While one might want to use w_c to minimize f , this is only possible for divisible applications. Here, we can checkpoint only at the end of a task, and the amount of work w can only be the sum of some task durations.

Consider a path starting after a checkpoint c_i (hence with a recovery r_i), and ending in a checkpoint c_j . The amount of computation w between these two checkpoints is necessarily of the form

$$W_{i,j}(k) = W_{i,j} + k.T, \text{ for some } k \in \mathbb{N},$$

where

- (i) T is the length of the iterations ($T = \sum_{\ell=0}^{n-1} t_\ell$), and
- (ii) $W_{i,j}$ the length between the end of task a_i and the end of task a_j (possibly of the next iteration), i.e.: $W_{i,j} = \sum_{\ell=i+1}^j t_\ell$ (case $j > i$), or $W_{i,j} = T - W_{j,i}$ (case $j < i$), or $W_{i,j} = T$ (case $j = i$).

Additionally, $k \mapsto W_{i,j}(k)$ is an increasing function, hence for all pairs (r_i, c_j) , there exists $k_{i,j}^*$ that minimizes the function $k \mapsto f(W_{i,j}(k), c_j, r_i)$. Let $W_{i,j}^* = W_{i,j}(k_{i,j}^*)$. Because $f(w, c_j, r_i)$ is decreasing for $w < w_{c_j}$, we have $W_{i,j}^* - T < w_{c_j}$ (otherwise $W_{i,j}^* - T$ would be a better solution). Finally, we denote

$$M^* = \max_i w_{c_i} + T.$$

Then, $M^* \geq \max_{i,j} W_{i,j}^*$, and by construction, we have the following property for M^* :

Lemma 5. *For all i, j, k_1, k_2 such that $W_{i,j} + k_1 \cdot T \geq W_{i,j} + k_2 \cdot T \geq M^*$,*

$$f(W_{i,j} + k_1 \cdot T, c_j, r_i) > f(W_{i,j} + k_2 \cdot T, c_j, r_i).$$

Finally, we let

$$k^* = \lfloor M^*/T \rfloor \quad (7)$$

denote the number of iterations that take place during time M^* . We are ready to bound the length between two successive checkpoints within an optimal pattern:

Lemma 6. *Given an optimal pattern, $(i_0, [m_1, m_2, \dots, m_{k_0}])$, then for all $1 \leq i \leq k_0$, $m_i - m_{i-1} \leq 2M^*$ (using $m_0 = 0$).*

Following the proof: In order to show this result, we show that if the length between two consecutive checkpoints was larger than the bound, then we could add an intermediate checkpoint and create a pattern of smaller slowdown.

Proof. We start by a preliminary property that we use in the following to show that if the length between two checkpoints is too high, then we can create a pattern of better slowdown by incorporating a checkpoint in the oversized interval.

Given a pattern $P = (i_0, [m_1, m_2, \dots, m_{k_0}])$, and given a transformation of this pattern into a pattern P' of equal length with an extra checkpoint of cost C (and recovery R) located between the $(i-1)^{\text{st}}$ and the i^{th} checkpoint of P , after W units of work, one can verify that

$$\begin{aligned} S(P) - S(P') &= \frac{\mathbb{E}_\lambda(W_i^P, C_i^P, R_{i-1}^P) - \mathbb{E}_\lambda(W, C, R_{i-1}^P) - \mathbb{E}_\lambda(W_i^P - W, C_i^P, R)}{\ell(P)} \\ &= \frac{W_i^P}{\ell(P)} f(W_i^P, C_i^P, R_{i-1}^P) - \frac{W}{\ell(P)} f(W, C, R_{i-1}^P) \\ &\quad - \frac{W_i^P - W}{\ell(P)} f(W_i^P - W, C_i^P, R) \end{aligned} \quad (8)$$

Indeed, $\ell(P) = \ell(P')$, and all inter-checkpoint intervals are identical (and have an equal cost) in P and P' , except for the interval inside which the extra checkpoint has been added.

We can now prove the result. We show the result by contradiction: assume there exists $i \leq k_0$ such that $m_i - m_{i-1} > 2M^*$. We denote by $i_0 = m_{i-1}[n]$ and $i_1 = m_i[n]$.

- Assume first that $c_{i_1} \geq c_{i_0}$. By monotony, $r_{i_1} \geq r_{i_0}$. We create the pattern P' such that we add to P an additional checkpoint after the task of type i_0 at the location $m_{i-1} + n \cdot k^*$ (which indeed corresponds to a task of type i_0). Then, using the properties of the slowdown function f , and because $W_i^P > T \cdot k^*$, we know that:

$$\begin{aligned} f(T \cdot k^*, C_{i-1}^P, R_{i-1}^P) &\leq f(T \cdot k^*, C_i^P, R_{i-1}^P) && \text{(growth)} \\ &< f(W_i^P, C_i^P, R_{i-1}^P) && \text{(shape of } f) \end{aligned}$$

Similarly, $W_i^P > W_i^P - T \cdot k^* \geq M^*$, then we have:

$$f(W_i^P - T \cdot k^*, C_i^P, R_{i-1}^P) \leq f(W_i^P, R_{i-1}^P, R_{i-1}^P)$$

Finally, plugging back these values into Equation (8), we obtain that P' has a better slowdown than P , contradicting the optimality.

- Assume now that $c_{i_1} \leq c_{i_0}$. By monotony, $r_{i_1} \leq r_{i_0}$. With a similar demonstration, we show that by including a checkpoint of size c_{i_1} at location $m_i - n \cdot k^*$ (which indeed corresponds to a task of type i_1), leads to the same result. \square

Theorem 5. *There exists an optimal pattern P whose length satisfies $\ell(P) \leq 2nM^*$, and which includes at most $2n^2(k^* + 1)$ tasks.*

Proof. From Lemma 2, we know that there exists an optimal pattern with at most n checkpoints. Using Lemma 6 which gives a bound on the inter-checkpoint time, we obtain the bound on the length. Thanks to Equation 7, we know that a length of M^* corresponds to at most $k^* + 1$ iterations (of n tasks each), which leads to the bound on the number of tasks. \square

We now need to check that k^* is polynomial in the size of the input. The size of the input is $O(n \max_i \log a_i)$, or equivalently $O(n \log T)$, because the n values a_i are encoded in binary. Here we make the natural assumption that $c_i = O(T)$ and $r_i = O(T)$ for $0 \leq i \leq n$, meaning that the largest checkpoint/recovery is not longer than a whole iteration². Recall that $w_{c_i} \approx \sqrt{\frac{2c_i}{\lambda}}$ [10, 47], hence $M^* = O(\sqrt{\frac{\max_i c_i}{\lambda}} + T)$ and $k^* = O\left(\frac{1}{\sqrt{\lambda T}} + 1\right)$. We obtain a polynomial value $k^* = O(n \log T)$ as soon

²Technically, we can relax the assumption to $c_i, r_i = O(T^n)$ without increasing the problem size.

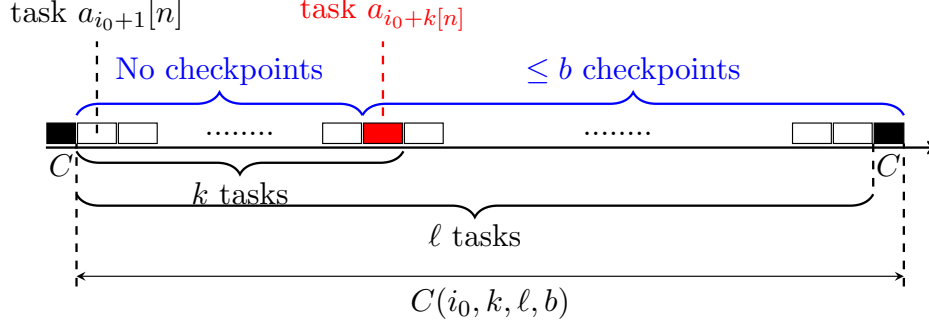


Figure 6: Illustration of the minimum expected execution time $C_{\min}(i_0, k, \ell, b)$ of a series of ℓ tasks as characterized by Lemma 7.

as $\mu = \frac{1}{\lambda} = O(T(n \log T)^2)$. This requires that the application MTBF is not too large in front of the iteration length, which makes full sense because otherwise we would not checkpoint more than very rarely, once every many iterations. In Section 4.4.2, we present a dynamic programming algorithm to compute an optimal pattern, whose complexity is polynomial in n and k^* . This complexity is indeed polynomial in the size of the instance under the very natural assumptions that we made.

4.4.2 Computing an optimal pattern

In the previous section we have shown the existence of an *optimal* pattern of polynomial length. Here, we show how one can compute an optimal pattern through a dynamic programming algorithm. This dynamic programming algorithm relies upon the previous results:

- We study patterns P of length at most $2nM^*$ (thanks to Theorem 5), and we know that an optimal pattern of this length contains a polynomial number of tasks;
- We consider different initial tasks in the pattern ($i_0 \in \{0, \dots, n-1\}$);
- We use the fact that there can be at most n checkpoints in the optimal pattern (thanks to Lemma 2).

The following lemma characterizes the minimal cost of a checkpoint path.

Lemma 7 (Minimum cost of a path). *The minimal expected execution time (or cost) of a checkpoint path (i) of length ℓ , (ii) whose first task is $a_{i_0+1}[n]$, (iii) with at most b checkpoints, (iv) where the $k-1$ first tasks are not*

checkpointed and, (v) where the last task is checkpointed, is given by;

$$\mathcal{C}_{\min}(i_0, k, \ell, b) = \min \begin{cases} \mathcal{C}_{\min}(i_0, k+1, \ell, b); \\ \mathbb{E}_{\lambda}(\sum_{i=1}^k w_{i_0+i[n]}, C_{i_0+k[n]}, R_{i_0}) \\ + \mathcal{C}_{\min}(i_0+k[n], 1, \ell-k, b-1) \end{cases} \quad (9)$$

when $\ell > 0$ and $b > 0$, and where we consider the following initialisation cases:

$$\begin{aligned} (a) \quad \mathcal{C}_{\min}(i_0, \ell+1, \ell, b) &= \begin{cases} 0 & \text{if } \ell = 0 \\ \infty & \text{otherwise} \end{cases} \\ (b) \quad \mathcal{C}(i_0, k, \ell > 0, 0) &= \infty. \end{aligned}$$

Please refer to Figure 6 for a graphical representation of $\mathcal{C}(i_0, k, \ell, b)$.

Proof. The result is showed recursively. We start with the initialisation cases. When no checkpoint is allowed ($b = 0$, case (b)), it is not even possible to checkpoint the last task (as required by condition (v)), so the cost is infinite. The case $k = \ell + 1$ leads to ℓ tasks not being checkpointed (condition (iv)), which contradicts the fact that the last task is checkpointed (condition (v)), except when the number of tasks is zero: in this case, we assume that no task is performed in this path and no checkpoint is taken.

We now move to the general case. Considering a path that verifies the condition of the lemma, we distinguish two cases:

- (i) The k^{th} task is not checkpointed, which leads to the first k tasks not being checkpointed, hence the minimum cost is $\mathcal{C}_{\min}(i_0, k+1, \ell, b)$;
- (i) The k^{th} task is checkpointed. The cost of the first part ($k-1$ tasks not checkpointed followed by this k^{th} task and its checkpoint is given by $\mathbb{E}_{\lambda}(\sum_{i=1}^k w_{i_0+i[n]}, C_{i_0+k[n]}, R_{i_0})$. The cost of the rest of the path is recursively expressed as the minimal cost of a path of length $\ell-k$ that starts after task $i_0+k[n]$ with $b-1$ checkpoints.

We then select the case that leads to the minimal expected execution time. \square

Thanks to Lemma 6, we know that in an optimal pattern, there are at most $2n(k^*+1)$ tasks between two checkpoints. So we can safely restrict our search space to $k = 1 \dots 2n(k^*+1)$ and consider that the cost for larger values of k is infinite. Hence, the previous recursive definition of the cost is applied to the design of the dynamic programming algorithm (Algorithm 1).

Theorem 6. $\text{PATTERN}(k^*, n)$ (Algorithm 1) returns the slowdown of the pattern of an optimal periodic schedule with time complexity $O((k^*)^2 n^5)$.

Proof. We use the fact that there exists an optimal periodic schedule whose pattern includes a number $m \times n$ of tasks with $m \leq 2n(k^*+1)$ and uses at

Algorithm 1 Finding the minimum slowdown of a pattern of size at most $2n^2(k^* + 1)$

```

1: procedure PATTERN( $k^*, n$ )
2:    $maxK \leftarrow 2n(k^* + 1)$ 
3:   for  $i_0 = 0$  to  $n - 1$  do                                      $\triangleright$  Initialization of ProgDyn
4:     for  $\ell = 0$  to  $2n^2(k^* + 1)$  do
5:       for  $b = 1$  to  $n$  do
6:         if  $\ell = 0$  then
7:            $C_{\min}(i_0, \ell + 1, \ell, b) \leftarrow 0$ 
8:         else
9:            $C_{\min}(i_0, \min(maxK, \ell + 1), \ell, b) \leftarrow \infty$ 
10:        for  $k = 1$  to  $\min(maxK, \ell)$  do
11:           $C_{\min}(i_0, k, \ell, 0) \leftarrow \infty$ 
12:        for  $i_0 = 0$  to  $n - 1$  do                                      $\triangleright$  Precompute  $\sum_{i=1}^k w_{i_0+i[n]}$ 
13:           $W[i_0, 1] \leftarrow t_{i_0+1[n]}$ 
14:          for  $k = 2$  to  $2n^2(k^* + 1)$  do
15:             $W[i_0, k] \leftarrow W[i_0, k - 1] + t_{i_0+k[n]}$ 
16:          for  $\ell = 1$  to  $2n^2(k^* + 1)$  do                                $\triangleright$  Computing the ProgDyn
17:            for  $i_0 = 0$  to  $n - 1$  do
18:              for  $k = \min(maxK - 1, \ell)$  downto 1 do
19:                for  $b = 1$  to  $n$  do
20:                   $C_{\min}(i_0, k, \ell, b) \leftarrow \min(C_{\min}(i_0, k + 1, \ell, b),$ 
21:                     $\mathbb{E}_\lambda(W[i_0, k], C_{i_0+k[n]}, R_{i_0}) + C_{\min}(i_0 + k[n], 1, \ell - k, b - 1))$ 
22:                   $SD = \infty$                                       $\triangleright$  Computing the minimal slowdown.
23:                   $T = \sum_{i=1}^n t_i$ 
24:                  for  $i_0 = 0$  to  $n - 1$  do
25:                    for  $m = 1$  to  $2n(k^* + 1)$  do
26:                       $SD_{\text{temp}} = C_{\min}(i_0, 1, mn, n)/mT$ 
27:                      if  $SD_{\text{temp}} < SD$  then
28:                         $SD \leftarrow SD_{\text{temp}}$ 
29:          return  $SD$ 

```

most n checkpoints (see Theorem 5). Algorithm 1 computes the minimum cost of all patterns including at most this number of tasks, then computes the minimum cost of a pattern whose number of tasks is a multiple of n . The slowdown we look for is indeed this cost. The complexity of the algorithm derives from the loop nest necessary to recursively compute \mathcal{C}_{\min} . \square

5 Simulation Results

In this section, we describe the experiments conducted to compare the proposed optimal checkpointing strategy with simpler heuristics. We perform simulations on three application scenarios: two from real-life applications, and one using synthetic parameters.

5.1 Experimental methodology

We detail here the algorithms used in the simulations and the various settings. All algorithms have been implemented in MATLAB and R. The corresponding code is publicly available at [12].

5.1.1 Application scenarios

We consider both real-life and synthetic application scenarios. For each scenario, we describe the parameters of the application: number of tasks in each iteration, computation cost of each task, checkpointing and recovery cost associated to each task.

Neuroscience application For the first application scenario, we extracted data from a representative neuroscience application, Spatially Localized Atlas Network Tiles (SLANT) [25]. The tasks composing an iteration of this application are described in Table 1, extracted from [17].

Table 1: Tasks of the neuroscience application.

Task	a_0	a_1	a_2	a_3	a_4	a_5	a_6
Execution time t_i	255	871	588	459	3050	804	1130
Checkpoint time C_i	22.22	61.11	33.33	50	283.33	16.67	61.11
Recovery time R_i	8.89	24.44	13.33	20	113.33	6.67	24.44

GCR application For the second application scenario, we consider a class of Krylov Subspace method GCR [14] solving the m -dimensional sparse linear system $A\mathbf{x} = \mathbf{b}$. Each iteration of the method is divided into n sub-iterations, whose computational and memory requirements increase from one sub-iteration to the next. The common way to control the number of iterative steps within an acceptable range is to adopt a restart strategy [36, 40, 46], that is, to fix a small value n (usually much less than m , such as 10,

20, etc.). If the last n -th sub-iteration does not lead to convergence, then the approximate solution \mathbf{x}^n is used as the initial value of a new iteration, and the GCR method is restarted. The process is repeated until a satisfactory approximate solution is found, as detailed in Algorithm 2.

Algorithm 2 GCR(n)

```

1:  $\mathbf{x}^0, \mathbf{r}^0 = A\mathbf{x}^0 - \mathbf{b}, \mathbf{p}^0 = P^{-1}\mathbf{r}^0, \mathbf{q}^0 = A\mathbf{p}^0$ 
2: for  $k = 1, 2, \dots$ , until convergence do
3:   for  $i = 0, 1, \dots, n - 1$  do
4:      $\beta = \frac{(\mathbf{r}^i, \mathbf{q}^i)}{(\mathbf{q}^i, \mathbf{q}^i)}$   $\triangleright 4m - 1$ 
5:      $\mathbf{x}^{i+1} = \mathbf{x}^i + \beta \mathbf{p}^i$   $\triangleright 2m$ 
6:      $\mathbf{r}^{i+1} = \mathbf{r}^i + \beta \mathbf{q}^i$   $\triangleright 2m$ 
7:     if  $\|\mathbf{r}^{i+1}\| \leq \varepsilon$  then
8:       exit
9:      $\mathbf{e} = P^{-1}\mathbf{r}^{i+1}$   $\triangleright 3m - 1$ 
10:     $\tilde{\mathbf{e}} = A\mathbf{e}$   $\triangleright 2nz(A) - 1$ 
11:    for  $l = 0, 1, \dots, i$  do
12:       $\alpha_l = \frac{(\tilde{\mathbf{e}}, \mathbf{q}^l)}{(\mathbf{q}^l, \mathbf{q}^l)}$   $\triangleright (i+1)(4m-1)$ 
13:       $\mathbf{p}^{i+1} = \mathbf{e} + \sum_{l=0}^i \alpha_l \mathbf{p}^l$   $\triangleright m + (i+1)m$ 
14:       $\mathbf{q}^{i+1} = \tilde{\mathbf{e}} + \sum_{l=0}^i \alpha_l \mathbf{q}^l$   $\triangleright m + (i+1)m$ 
15:     $[\mathbf{x}^0, \mathbf{r}^0, \mathbf{p}^0, \mathbf{q}^0] \leftarrow [\mathbf{x}^n, \mathbf{r}^n, \mathbf{p}^n, \mathbf{q}^n]$ 

```

Each iteration of the outer loop k has n tasks ($i = 0, 1, \dots, n - 1$), each task corresponding to one sub-iteration of the loop on i . The number of non-zero elements of sparse matrix A is denoted as $nz(A)$. We assume that $m = 100000$, $nz(A) = 27m$, and the preconditioner matrix P is a diagonal matrix in the simulation. We pick (somewhat arbitrarily) 27 because it is the size of a $3 \times 3 \times 3$ cube for a neighborhood of interactions, so the matrix has 27 diagonals (3D-stencil for Jacobi or Gauss-Seidel, typically). We computed that the number of floating-point operations for task i is $f_i = (6m - 1)i + 19m - 4 + 2nz(A)$, see Table 2. The vectors that need to be saved if we checkpoint after task i and the corresponding size of the checkpoint are also detailed in this table.

We consider here that the computing platform has unit speed $s = 1$, so that $t_i = f_i/s = f_i$. In order to test different scenarios for the relative cost of checkpoint compared to computations, we define the Communication-to-Computation Ratio (CCR) as ratio between the cost of communicating one byte to the cost of computing one flop. With the choice $s = 1$, the CCR is exactly the inverse of the bandwidth. Hence, from the size of the memory to checkpoint M_i , we compute the time for a checkpoint: $C_i = M_i \times \text{CCR}$.

We conducted experiments with $\text{CCR} = 2$ and $\text{CCR} = 5$.

Task	Floating point operations f_i	Vectors to checkpoint	Memory size M_i
a_0	$19m - 4 + 2nz(A)$	$\mathbf{p}^0, \mathbf{p}^1, \mathbf{q}^0, \mathbf{q}^1, \mathbf{r}^1, \mathbf{x}^1$	$6m$
a_1	$(6m - 1) + 19m - 4 + 2nz(A)$	$\mathbf{p}^0, \mathbf{p}^1, \mathbf{p}^2, \mathbf{q}^0, \mathbf{q}^1, \mathbf{q}^2, \mathbf{r}^2, \mathbf{x}^2$	$8m$
\dots	\dots	\dots	\dots
a_{n-2}	$(6m - 1)(n - 2) + 19m - 4 + 2nz(A)$	$\mathbf{p}^0, \dots, \mathbf{p}^{n-1}, \mathbf{q}^0, \dots, \mathbf{q}^{n-1}, \mathbf{r}^{n-1}, \mathbf{x}^{n-1}$	$(2n + 2)m$
a_{n-1}	$(6m - 1)(n - 1) + 19m - 4 + 2nz(A)$	$\mathbf{p}^0, \mathbf{q}^0, \mathbf{r}^0, \mathbf{x}^0$	$4m$

Table 2: Tasks composing the GCR application.

Synthetic application The third application scenario is randomly generated. We consider an iterative application composed of $N = 10^3$ iterations, each iteration has $n = 10$ or 20 cyclic tasks. We assume that the execution time t_i of each task a_i follows a probability distribution \mathcal{D} , where \mathcal{D} is $\text{UNIFORM}(a, b)$. The default instantiation for this distribution is $\mu_{\mathcal{D}} = 550$ for $\text{UNIFORM}[100, 1000]$.

In this third application scenario, we use two strategies to choose the checkpoint time. Either we set checkpoint times as $C_i = \eta t_i$, where η is the proportion of checkpoint time to the execution time of each task; or we set checkpoint times to be taken in $\text{UNIFORM}[10, 100]$, independently of the task running time. We use $R = C$ for the recovery time and a fixed downtime $D = 5$. For the first choice of checkpoint times, we conducted experiments with $\eta = 0.1$.

5.1.2 Failure scenarios

We consider different failure rates. To allow for consistent comparisons of results across different iterative processes, we fix the probability that a failure occurs during each iteration, which we denote as p_{fail} , and then simulate the corresponding failure rate. Formally, for a given p_{fail} value, we compute the failure rate λ such that $p_{\text{fail}} = 1 - e^{-\lambda T}$, where T is the execution time per iteration with n tasks. We conduct experiments for six p_{fail} values: $10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{-0.5}$ and $10^{-0.1}$. For example, $p_{\text{fail}} = 10^{-2}$ means one failure will occur every 100 iterations on average.

For each experiment, the simulations are performed on 100 randomly generated instances $\{\mathcal{I}_1, \dots, \mathcal{I}_{100}\}$. For all i , an instance \mathcal{I}_i is a pair $(\mathcal{S}_i, \mathcal{F}_i)$, where \mathcal{S}_i (resp. \mathcal{F}_i) is the application (resp. failure) scenario associated to the instance. For both real-life scenarios, \mathcal{S}_i corresponds to the values presented in the previous tables, while for the synthetic application scenario, \mathcal{S}_i is randomly generated as described above.

5.1.3 Optimal strategy

The algorithm used to simulate the optimal strategy is detailed in Algorithm 1. For an instance \mathcal{I} , we define $MS_{opt}(p_{fail})(\mathcal{I})$ to be the makespan corresponding to the failure rate p_{fail} .

5.1.4 Reference strategies

We consider three reference *greedy* strategies:

1. CKPTEACHITER, consists in checkpointing at the end of each iteration, that is, a checkpoint is taken after the last task a_{n-1} of each iteration. For an instance \mathcal{I} , we define $MS_{iter}(p_{fail})(\mathcal{I})$ to be the makespan corresponding to the failure rate p_{fail} .
2. CKPTEACHTASK, consists in checkpointing after every task a_i of every iteration. For an instance \mathcal{I} , we define $MS_{task}(p_{fail})(\mathcal{I})$ to be the makespan corresponding to the failure rate p_{fail} .
3. CKPTYD, uses an adaption of the classical Young-Daly formula to compute the optimal checkpoint size in the case of divisible application (where one can checkpoint at any time-step). First it chooses the task of an iteration with minimum checkpoint size C_{min} . Only the result of this task will (possibly) be checkpointed, and we use the Young-Daly formula to compute how many iterations to include in between two checkpoints. More specifically, the number of iterations in this pattern is set to

$$\mathcal{P}_{YD} = \max \left(1, \text{round} \left(\frac{1}{T} \sqrt{\frac{2C_{min}}{\lambda}} \right) \right).$$

For an instance \mathcal{I} , we define $MS_{YD}(p_{fail})(\mathcal{I})$ to be the makespan corresponding to the failure rate p_{fail} .

5.2 Results

Results of the neuroscience application are reported in Figure 7. For the GCR application, the results are reported in Figures 8a and 8b, and the results of the synthetic application scenario are reported in Figures 9a and 9b. In these figures, we normalize the results to the makespan obtained by the optimal strategy. Namely, for each problem instance, the red (leftmost) boxplots represent the ratio $\frac{MS_{iter}(p_{fail})(\mathcal{I})}{MS_{opt}(p_{fail})(\mathcal{I})}$ for different values of p_{fail} and different scenarios, while the green (middle) boxplots represent the ratio $\frac{MS_{task}(p_{fail})(\mathcal{I})}{MS_{opt}(p_{fail})(\mathcal{I})}$; then the blue (rightmost) boxplots represent the ratio $\frac{MS_{YD}(p_{fail})(\mathcal{I})}{MS_{opt}(p_{fail})(\mathcal{I})}$.

General comments As expected, the greedy heuristics do not perform as well as the optimal strategy derived in this work. We now study and explain the behaviors of the different heuristics.

Checkpointing after each task, as done by CKPTEACHTASK, gives worst performance when p_{fail} is small (very few failures). Its performance improves significantly when the number of failures increases. This behavior is expected as it is a consequence of the very high number of checkpoints that are taken.

On the contrary, the Young-Daly inspired heuristic (CKPTYD) gives almost optimal results when there are very few failures, and its performance gets very poor when the number of failures increases. Again, this behavior is expected, with very few failures, if the frequency of checkpointing is of the same order of magnitude as in the optimal solution, the fact that the checkpointing decision that is taken is not optimal has little impact, because the checkpoint overhead is very low. With numerous failures, then this strategy, which is limited to at most one checkpoint per iteration, does not checkpoint often enough, and the loss in work when there is a failure gets too expensive.

Finally CKPTEACHITER is probably the less interesting strategy as its performance is always worse than CKPTYD. Its performance improves then gets worse when p_{fail} increases: when p_{fail} is very small, (i) it does not choose the task with smallest checkpoint size and (ii) it checkpoints too often compared to CKPTYD which would allow to checkpoint after several iterations and not just one; conversely, when p_{fail} is very large, checkpointing once after each iteration is not enough, thus the relative cost of the CKPTEACHITER strategy increases. It is still interesting to see that the difference with CKPTYD remains always small, contrarily to what one could have expected, therefore hinting that the importance of finding the smallest checkpoint size is not critical, and mitigating the impact of checkpointing too often when there are too few failures (although one checkpoint after each task, as in CKPTEACHTASK, is too often).

For all results, there exists an intermediary phase: when the frequency of failures ranges from a few iterations per failure to a few failures per iteration, then all greedy heuristic perform poorly, and the optimal solution provides significant gains.

Application specific comments For the neuroscience application, in Table 3, we also report the number of tasks given by the upper bound of Theorem 5, and the number of tasks actually occurring in the optimal pattern. As expected, the latter is smaller than the former, but we observe that the difference is quite significative: the bound of Theorem 5 is overly pessimistic. In addition, we see that both numbers decrease when p_{fail} becomes larger.

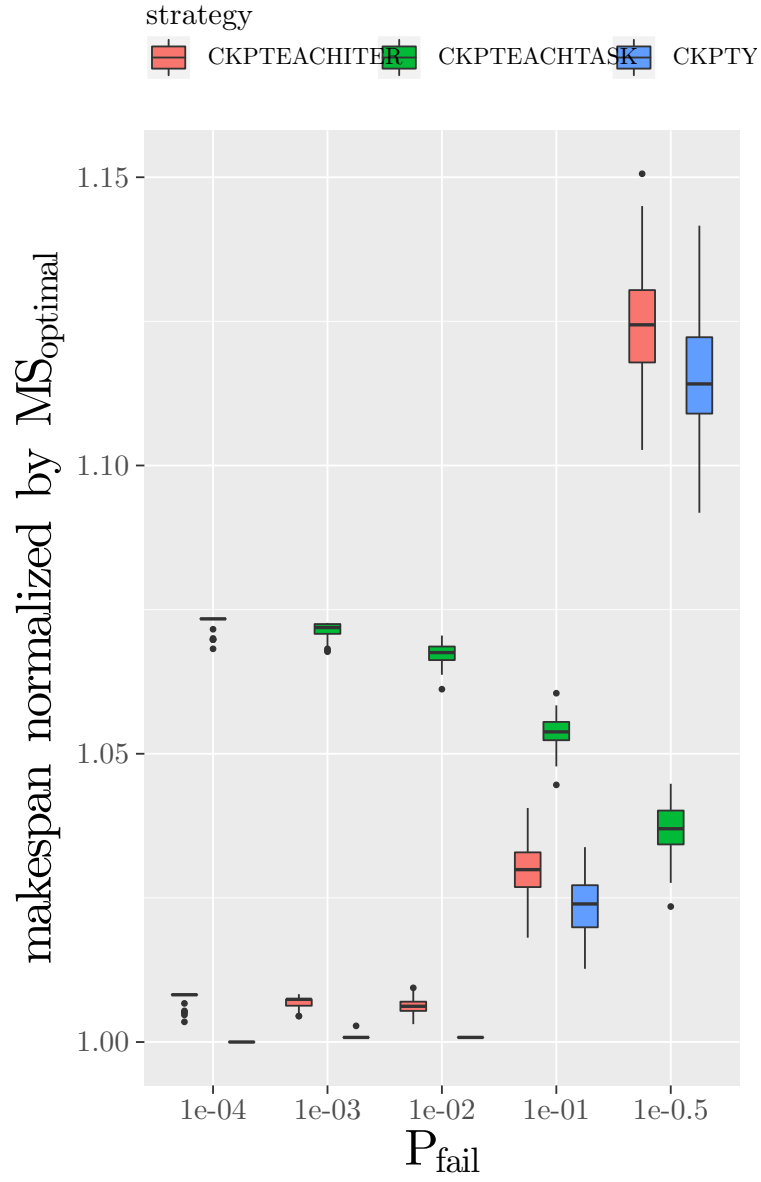


Figure 7: The boxplots represent the performance of the checkpoint strategies with different failure probabilities for the neuroscience application.

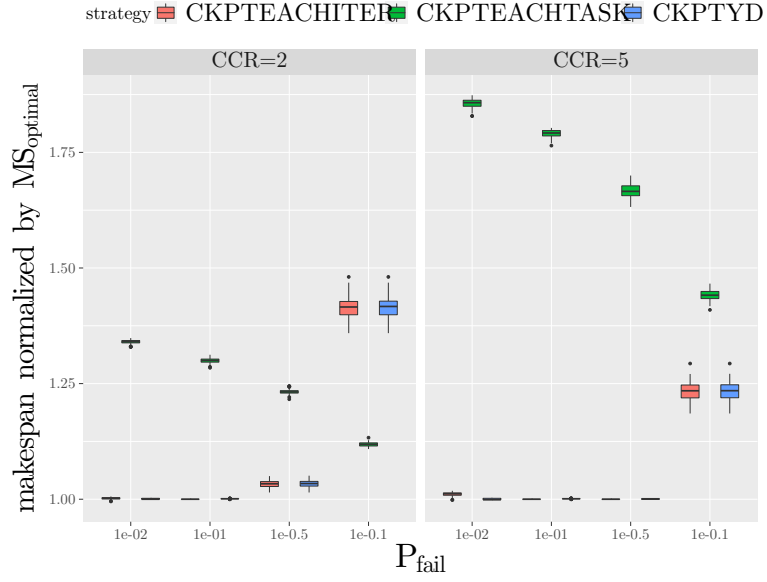
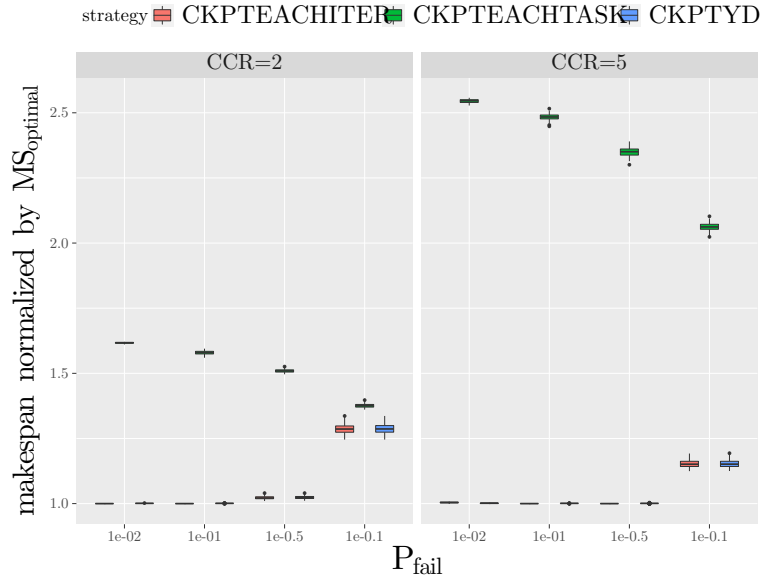
(a) $n = 10$ (b) $n = 20$ Figure 8: The boxplots represent the performance of the checkpoint strategies with different failure probabilities for $\text{GCR}(n)$.

Table 3: Task number from the bound and in the optimal pattern for the neuroscience application.

p_{fail}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	$10^{-0.5}$
Bound	2940	980	392	196	196
Optimal pattern	49	14	7	7	7

For the GCR application (see Figures 8a and 8b), when CCR goes from 2 to 5, CKPTEACHITER and CKPTYD are further away from the optimal strategy when p_{fail} is small (for 10^{-2} and 10^{-1}), while both strategies are closer to the optimal strategy when p_{fail} is large (for $10^{-0.5}$ and $10^{-0.1}$), and CKPTEACHTASK is further away from the optimal strategy for all p_{fail} values. In addition, when n goes from 10 to 20, CKPTEACHITER and CKPTYD are closer to the optimal strategy, while the CKPTEACHTASK is further away from the optimal strategy (up to 200% worse).

For the synthetic application (see Figures 9a and 9b), when n goes from 10 to 20, CKPTEACHITER and CKPTYD are closer to the optimal strategy when p_{fail} is small (for 10^{-3} and 10^{-2}), while both strategies are further away from the optimal strategy when p_{fail} is large (for 10^{-1} and $10^{-0.5}$), and CKPTEACHTASK is further away from the optimal strategy for all p_{fail} values. In addition, when the checkpoint time becomes independent from the running time, CKPTEACHITER and CKPTYD hardly change, while the variance of CKPTEACHTASK becomes larger.

An important factor that influences the performance of checkpointing strategies, and more precisely of the checkpointing and recovery overheads, is the data-intensiveness of the application. We discuss its impact on the performance of checkpointing strategies in two ways. For the GCR application, in order to test the impact of different checkpoint costs on the strategies, we vary the checkpoint cost by varying the CCR value. As expected, CKPTEACHTASK is further away from the optimal strategy for all p_{fail} values, since checkpoints become more expensive. Also, CKPTEACHITER and CKPTYD are further away from the optimal strategy when p_{fail} is small because checkpoints become more expensive, while both strategies are closer to the optimal strategy when p_{fail} is large (since the probability of failures occurring at checkpoint generation and recovery increases). For the synthetic application, in order to test the impact of the correlation between checkpoint costs and task running times on the strategies, we let the checkpoint time move from dependent to independent of the task running time. Then, the variance of CKPTEACHTASK becomes larger since the checkpoints set after each task all change. On the contrary, CKPTEACHITER and CKPTYD hardly change.

In summary, no greedy heuristic is able to give close-to-optimal

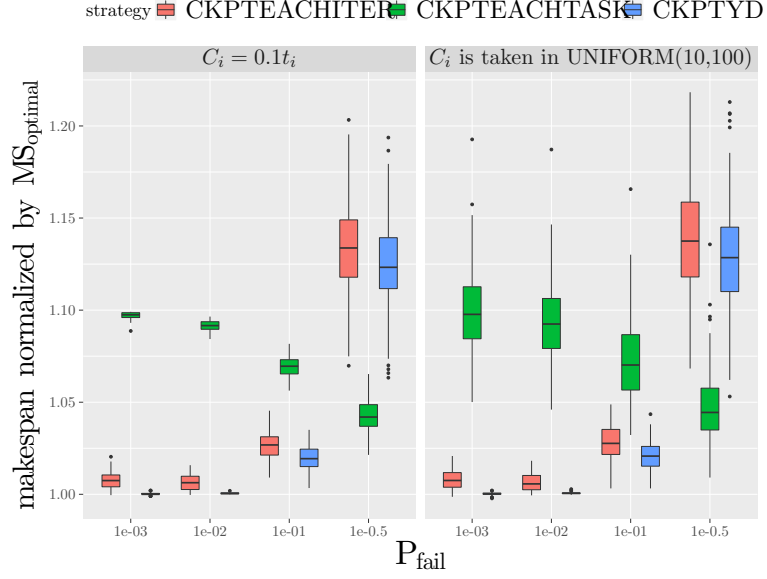
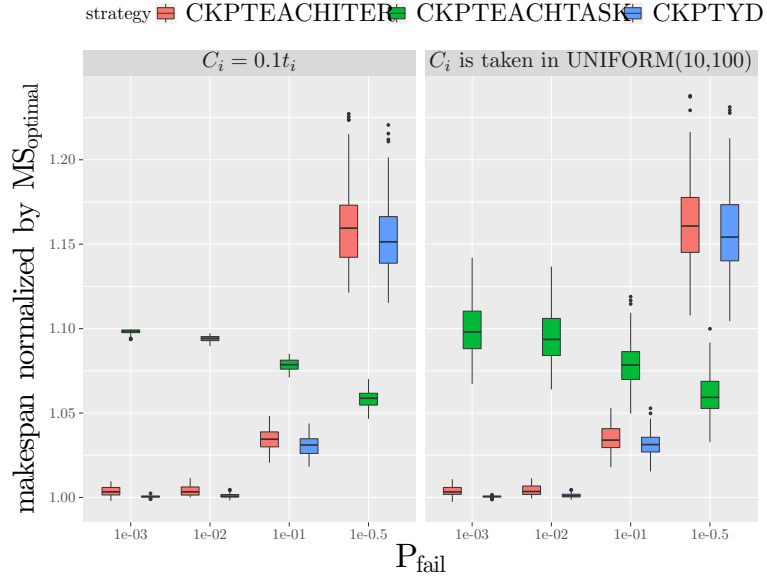
(a) $n = 10$ (b) $n = 20$

Figure 9: The boxplots represent the performance of the checkpoint strategies with different failure probabilities for the synthetic application.

makespan for every value of p_{fail} : CKPTYD is better with very few failures, while CKPTEACHTASK is better when there are many failures. For these extreme scenarios, using the ad-hoc greedy heuristic is a good solution to trade-off the complexity of finding the solution with the gain in performance. However, in intermediary scenarios, the best of the greedy heuristics can still increase the time to solution by 25%, showing the importance of computing the correct solution! Altogether, our proposed optimal scheme enables us to carefully optimize the checkpoint pattern for all problem instances.

6 Conclusion

In this work, we have investigated checkpointing strategies for iterative applications. Each iteration is composed of a chain of tasks, and these tasks have different lengths and different checkpoint costs. Simple approaches would checkpoint either every task, or the last task at the end of each iteration. An approach inspired by the Young/Daly formula would select the task with lowest checkpoint cost and checkpoint every p^{th} instance of that task, where p is computed so that the period length obeys the formula approximately. But what is the optimal strategy? The main contributions of this paper are threefold: (i) we have shown that there exists a periodic strategy that is optimal; (ii) we have provided a dynamic-programming algorithm that computes the optimal period; and (iii) we have shown through a set of experiments that the gains over the other approaches are significant, and that the optimal strategy is the only one achieving a robust solution for all problem instances cases. Given the importance of iterative applications in HPC, we expect these contributions to greatly improve the deployment of resilient solutions at scale.

Future work will be devoted to dealing with iterative applications whose iterations are composed of a Directed Acyclic Graph (DAG) of tasks, not just a linear chain. Such applications are ubiquitous in real-time systems. However, the mere fact that several tasks may execute concurrently on the platform raises very complicated challenges [21, 22], and most likely only heuristic (suboptimal) algorithms will be obtained.

References

- [1] Emmanuel Agullo, Luc Giraud, Abdou Guermouche, Jean Roman, and Mawussi Zounon, *Towards resilient parallel linear krylov solvers: recover-restart strategies* (2013).
- [2] ———, *Numerical recovery strategies for parallel resilient krylov linear solvers*, Numerical Linear Algebra with Applications **23** (2016), no. 5, 888–905.

- [3] Guillaume Aupy, Anne Benoit, Henri Casanova, and Yves Robert, *Checkpointing strategies for scheduling computational workflows*, International Journal of Networking and Computing **6** (2016), no. 1, 2–26.
- [4] Marin Bougeret, Henri Casanova, Mikael Rabie, Yves Robert, and Frédéric Vivien, *Checkpointing strategies for parallel jobs*, Technical Report 7520, INRIA, France, 2011. Short version appears in SC’2011.
- [5] Patrick G Bridges, Kurt B Ferreira, Michael A Heroux, and Mark Hoemmen, *Fault-tolerant linear solvers via selective reliability*, arXiv preprint arXiv:1206.1390 (2012).
- [6] Chandra Chekuri, Waqar Hasan, and Rajeev Motwani, *Scheduling problems in parallel query optimization*, Pods’1995, the 14th acm sigact-sigmod-sigart symposium on principles of database systems, 1995, pp. 255–265.
- [7] Zizhong Chen, *Algorithm-based recovery for iterative methods without checkpointing*, Proceedings of the 20th international symposium on high performance distributed computing, 2011, pp. 73–84.
- [8] ———, *Online-abft: An online algorithm based fault tolerance scheme for soft error detection in iterative methods*, ACM SIGPLAN Notices **48** (2013), no. 8, 167–176.
- [9] Alok Choudhary, Wei-keng Liao, Donald Weiner, Pramod Varshney, Richard Linderman, Mark Linderman, and Russell Brown, *Design, implementation and evaluation of parallel pipelined STAP on parallel computers*, IEEE Transactions on Aerospace and Electronic Systems **36** (April 2000), no. 2, 655–662.
- [10] J. T. Daly, *A higher order estimate of the optimum checkpoint interval for restart dumps*, Future Generation Comp. Syst. **22** (2006), no. 3, 303–312.
- [11] Ewa Deelman, James Blythe, Yolanda Gil, and Carl Kesselman, *Grid resource management*, 2003.
- [12] Yishu Du, *Code for simulations*, 2020. <https://github.com/Yishu0604/Optimal-Checkpointing-Strategies-for-Iterative-Applications>.
- [13] Yishu Du, Loris Marchal, Guillaume Pallez, and Yves Robert, *Robustness of the Young/Daly formula for stochastic iterative applications*, ICPP’2020, 2020.
- [14] Stanley C Eisenstat, Howard C Elman, and Martin H Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM Journal on Numerical Analysis **20** (1983), no. 2, 345–357.
- [15] James Elliott, Mark Hoemmen, and Frank Mueller, *Evaluating the impact of sdc on the gmres iterative solver*, 2014 ieee 28th international parallel and distributed processing symposium, 2014, pp. 1193–1202.
- [16] Stanley P Frankel, *Convergence rates of iterative treatments of partial differential equations*, Mathematical Tables and Other Aids to Computation **4** (1950), no. 30, 65–75.
- [17] Ana Gainaru, Brice Goglin, Valentin Honoré, and Guillaume Pallez, *Profiles of upcoming HPC Applications and their Impact on Reservation Strategies*, Technical Report RR-9359, Inria & Labri, Univ. Bordeaux, 2020.
- [18] Fernando Guirado, Ana Ripoll, Concepció Roig, Aura Hernandez, and Emilio Luque, *Exploiting throughput for pipeline execution in streaming image processing applications*, Euro-par 2006, parallel processing, 2006, pp. 1095–1105.
- [19] Fernando Guirado, Ana Ripoll, Concepcio Roig, and Emilio Luque, *Optimizing latency under throughput requirements for streaming applications on cluster execution*, Cluster computing, 2005. ieee international, September 2005, pp. 1–10.
- [20] Martin H Gutknecht, *Variants of bicgstab for matrices with complex spectrum*, SIAM journal on scientific computing **14** (1993), no. 5, 1020–1033.

- [21] Li Han, Louis-Claude Canon, Henri Casanova, Yves Robert, and Frédéric Vivien, *Checkpointing workflows for fail-stop errors*, IEEE Trans. Computers **67** (2018), no. 8, 1105–1120.
- [22] Li Han, Valentin Le Fèvre, Louis-Claude Canon, Yves Robert, and Frédéric Vivien, *A Generic Approach to Scheduling and Checkpointing Workflows*, Int. Journal of High Performance Computing Applications **33** (2019), no. 6, 1255–1274.
- [23] Timothy D. R. Hartley, Ahmed R. Fasih, Charles A. Berdanier, Fusun Ozguner, and Ümit V. Çatalyürek, *Investigating the Use of GPU-Accelerated Nodes for SAR Image Formation*, Proceedings of the IEEE international conference on cluster computing, workshop on parallel programming on accelerator clusters (ppac), 2009.
- [24] Thomas Herault and Yves Robert (eds.), *Fault-Tolerance Techniques for High-Performance Computing*, Computer Communications and Networks, Springer Verlag, 2015.
- [25] Yuankai Huo, Zhoubing Xu, Yunxi Xiong, Katherine Aboud, Prasanna Parvathaneni, Shunxing Bao, Camilo Bermudez, Susan M Resnick, Laurie E Cutting, and Bennett A Landman, *3D whole brain segmentation using spatially localized atlas network tiles*, NeuroImage **194** (2019), 105–119.
- [26] Jihie Kim, Yolanda Gil, and Marc Spraragen, *A knowledge-based approach to interactive workflow composition*, 14th international conference on automatic planning and scheduling (icaps 04), 2004.
- [27] Kathleen Knobe, James M. Rehg, Arun Chauhan, Rishiyur S. Nikhil, and Umakishore Ramachandran, *Scheduling constrained dynamic applications on clusters*, Supercomputing'1999, the 1999 acm/IEEE conference on supercomputing, 1999, pp. 46.
- [28] Julien Langou, Zizhong Chen, George Bosilca, and Jack Dongarra, *Recovery patterns for iterative methods in a parallel unstable environment*, SIAM Journal on Scientific Computing **30** (2008), no. 1, 102–116.
- [29] William L Oberkampf and Christopher J Roy, *Verification and validation in scientific computing*, Cambridge University Press, 2010.
- [30] Muhammed Emin Ozturk, Gagan Agrawal, Yukun Li, and Ching-Shan Chou, *Handling soft errors in krylov subspace methods by exploiting their numerical properties*, EasyChair, 2020.
- [31] Carlos Pachajoa, Markus Levonyak, and Wilfried N Gansterer, *Extending and evaluating fault-tolerant preconditioned conjugate gradient methods*, 2018 IEEE/ACM 8th workshop on fault tolerance for HPC at extreme scale (FTXS), 2018, pp. 49–58.
- [32] Carlos Pachajoa, Markus Levonyak, Wilfried N Gansterer, and Jesper Larsson Träff, *How to make the preconditioned conjugate gradient method resilient against multiple node failures*, Proceedings of the 48th international conference on parallel processing, 2019, pp. 1–10.
- [33] Carlos Pachajoa, Christina Pacher, Markus Levonyak, and Wilfried N Gansterer, *Algorithm-based checkpoint-recovery for the conjugate gradient method*, 49th international conference on parallel processing-ICPP, 2020, pp. 1–11.
- [34] Anthony Rowe, Dimitrios Kalaitzopoulos, Michelle Osmond, Moustafa Ghanem, and Yike Guo, *The discovery net system for high throughput bioinformatics*, Bioinformatics **19** (2003), no. Suppl 1, i225–31.
- [35] Christopher J Roy and William L Oberkampf, *A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing*, Computer methods in applied mechanics and engineering **200** (2011), no. 25–28, 2131–2144.
- [36] Y. Saad, *Iterative methods for sparse linear systems*, 2nd ed., Society for Industrial and Applied Mathematics, 2003.

- [37] Youcef Saad and Martin H Schultz, *Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on scientific and statistical computing **7** (1986), no. 3, 856–869.
- [38] Piyush Sao and Richard Vuduc, *Self-stabilizing iterative solvers*, Proceedings of the workshop on latest advances in scalable algorithms for large-scale systems, 2013, pp. 1–8.
- [39] Olcay Sertel, Jun Kong, Hiroyuki Shimada, Ümit V. Çatalyürek, Joel H. Saltz, and Metin N. Gurcan, *Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development*, Pattern Recognition **42** (2009), no. 6, 1093–1103.
- [40] Valeria Simoncini, *On the convergence of restarted krylov subspace methods*, SIAM Journal on Matrix Analysis and Applications **22** (2000), no. 2, 430–452.
- [41] Dingwen Tao, Sheng Di, Xin Liang, Zizhong Chen, and Franck Cappello, *Improving performance of iterative methods by lossy checkpointing*, Proceedings of the 27th international symposium on high-performance parallel and distributed computing, 2018, pp. 52–65.
- [42] Dingwen Tao, Shuaiwen Leon Song, Sriram Krishnamoorthy, Panruo Wu, Xin Liang, Eddy Z Zhang, Darren Kerbyson, and Zizhong Chen, *New-sum: A novel online abft scheme for general iterative methods*, Proceedings of the 25th acm international symposium on high-performance parallel and distributed computing, 2016, pp. 43–55.
- [43] Top500, *Top 500 Supercomputer Sites*, 2018. <https://www.top500.org/lists/2018/11/>.
- [44] Sam Toueg and Özalp Babaoğlu, *On the optimum checkpoint selection problem*, SIAM J. Comput. **13** (1984), no. 3.
- [45] David Young, *Iterative methods for solving partial difference equations of elliptic type*, Transactions of the American Mathematical Society **76** (1954), no. 1, 92–111.
- [46] David M Young and Kang C Jea, *Generalized conjugate-gradient acceleration of non-symmetrizable iterative methods*, Linear Algebra and its applications **34** (1980), 159–194.
- [47] John W. Young, *A first order approximation to the optimum checkpoint interval*, Comm. of the ACM **17** (1974), no. 9, 530–531.



**RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES**

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399